METHOD OF SEARCH CONTENT ENHANCEMENT

Related Applications

The contents of the following listed applications are hereby incorporated by reference:

(1) U.S. Patent application, serial # 10/157,243, filed on 05/30/2002 and entitled "Method and Apparatus for Providing Multiple Views of Virtual Documents."

(2) U.S. Patent application, serial # 10/159,373, filed on 06/03/2002 and entitled "A System and Method for Generating and Retrieving Different Document Layouts from a Given Content."

(3) U.S. Patent application, serial # 10/180,195, filed on 06/27/2002 and entitled "Retrieving Matching Documents by Queries in Any National Language."

(4) U.S. Patent application, (YOR920020141), filed on 07/23/2002 and entitled "Method of Search Optimization Based on Generation of Context Focused Queries."

(5) U.S. Patent application, serial # 10/209,619 filed on 07/31/2002 and entitled "A Method of Query Routing Optimization."

(6) U. S. Patent application, serial # 10/066,346 filed on 02/01/2002 and entitled "Method and System for Searching a Multi-Lingual Database."

(7) U.S. Patent application, serial #10/229,552 filed on 8/28/2002 and entitled "Universal Search Management Over One or More Networks."

(8) U.S. Patent application, serial #10/180,195 filed on 6/26/2002 and entitled "An International Information Search and Delivery System Providing Search Results Personalized to a Particular Natural Language."

(9) U.S. Patent application, serial # (CHA920030010US1) filed on even date herewith entitled "Method of Self Enhancement of Search Results Through Analysis Of System Logs"

Field of the Invention

The present invention relates to performing keyword searches and obtaining search results on database networks. More particularly, it relates to the improvement of the effectiveness of searches in obtaining desired search results.

Background of the Invention

Internet text retrieval systems accept a statement for requested information in terms of a search query S made up of a plurality of keywords $T_1, T_2, \ldots T_i, \ldots T_n$ and return a list of documents that contain matches for the search query terms. To facilitate the performance of such searches on internet databases, search engines have been developed that provide a query interface to the information containing sources and return search results ranked sequentially on how well the listed documents match the search query. The effectiveness in obtaining desired results varies from search engine to search engine. This is particularly true in searching certain product support databases which can be heavily weighted with technical content and the queries tend to be repetitive. In such databases, information can be in a number of natural languages, both in analog and digital form, and in a number of different formats, and in multiple machine languages. The relevancy of the search results depends on many factors, one being on the specificity of the search query. If the search query was specific enough, the probability of getting relevant results is generally higher. For example, the probability of getting documents on 'Java exception handling' is higher for the query 'Java exception' than for the query 'exception'. At the same time, some relevant documents may be excluded by a specific search query, because the query does not contain certain combinations of terms, contains superfluous terms or address the same subject matter using different words. For instance, as shown in Figure 1, if the query is 'video player for PC', the search engine may not be able to find and return relevant documents that are not about personal computers and/or instead of using 'video player' contain terms like 'DVD driver' or 'multimedia software'. Approaches to broaden searches by adding synonymous search terms and disregarding bad query terms are known. However, results using these known approaches have not been entirely satisfactory in turning up relevant documents and/or require additional real time examination of database logs and/or databases.

Therefore it is an object of the present invention to provide an improvement in search engine search results.

Another object of the present invention is to broaden search results to uncover relevant documents that do not contain requested query terms.

It is further an object of the present invention to provide requested information to searchers in selected technical areas.

Brief Description of the Invention

Whenever a document is going to be included into the textual database, a semantic binder is used to associate the document with one or more semantic nodes which are defined in a semantic taxonomy. When a search is performed, a search application looks through a semantic dictionary (which contains a table mapping queries to nodes on the semantic taxonomy) to see whether any corresponding semantic node can be found for the searchers query. If a match is found, the search application transforms the user's query into [”original query” OR “semantic node”] so that relevant documents, even they do not contain any user's keyword, can also be found in the database. The system binds semantic nodes arranged in a hierarchical structure of the taxonomy using a Log Analyzer which periodically looks through the system log for new queries and through textual indices for documents added to the database to generate the semantic dictionary and to bind the semantic nodes to the queries in the textual indices of the documents.

Since the above analysis arrangement is performed on on all customer queries, the search system enhancements have a direct effect on customer satisfaction. Further because the query log analysis and relevant document identification is performed off-line, response time to customer queries is not affected. In addition, with the search enhancements of the present invention the search system learns from user iterations.

Description of the Drawings

Figure 1 is a schematic diagram illustrating limitations in a prior art search process;

Figure 2 is a schematic diagram for system organization of an on-line area network;

Figure 3 is a schematic diagram of a private network incorporating the present invention and connected to the network shown in Figure 2;

Figure 4 is a schematic diagram showing a search system using the document semantic taxonomy of the present invention;

Figure 5 is a schematic diagram showing the generation of a semantic dictionary and modification of the document textual indices for the present invention;

Figures 6 and 7 are schematic flow diagrams showing operation of the textual analyzer of Figure 5;

Figure 8 is an overall operational schematic diagram showing the operation of the system of Figures 4 and 5; and

Figure 9 is a more detailed schematic diagram of Figure 5.

Detailed Description of the Invention

Referring now to Figure 2, communication between a plurality of user computers 100a to 100n and a plurality of information servers 102a to 102n is accomplished via an on-line service through a wide area network such as the Internet 104 that includes network node servers. The network node servers manage network traffic such as the communications between any given user's computer and an information server.

The computers 100 are equipped with communications software, including a WWW browser such as the Netscape browser of Netscape Communications Corporation, that allows a shopper to connect and use on-line shopping services via the Internet. The software on a user's computer 100 manages the display of information received from the servers to the user and communicates the user's actions back to the appropriate information servers 102 so that additional display information may be presented to the user or the information acted on. The connections 106 to the network nodes of the Internet may be established via a modem or other means such as a cable connection.

The servers illustrated in Figure 2, and discussed hereafter, are those of merchants which, for a fee provide products, services and information over the Internet. While the following discussion is directed at communication between shoppers and such merchants over the Internet, it is generally applicable to any information seeker and any information provider on a network. (For instance, the information provider can be a library such as a University library, a public library or the Library of Congress or other type of information providers.) Information regarding a merchant and the merchant's products is stored in a shopping database 108 to which the merchants servers 102 have access. This may be the merchants own database or a database of a supplier of the merchant. All product information accessible by the merchant servers that is publishable as web pages is indexed and a full-text index database 110 which records the number of occurrences of each of the words and their use in the location. In addition to the servers of individual merchants, and other information providers, there are the servers 114a to 114 of plurality of search service providers, such as Google of Google, Inc., which providers maintain full text indexes 116 of the products of the individual merchants 102a to 102n obtained by interrogating the product information databases 108 of the individual merchants. Some of these search service providers, like Google, are general purpose search providers while others are topic specific search providers.

The merchants and the search application service providers each may maintain a database of information about shoppers and their buying habits to customize on-line shopping for the shopper. Operations to accomplish a customized electronic shopping environment for the shopper include accumulating data regarding the shopper's preferences. Data relating to the electronic shopping options, such as specific sites and specific products selected by the shopper, entry and exit times for the sites, number of visits to the sites, etc., are recorded and processed by each merchant to create a shopping profile for the shopper. Raw data may then be processed to create a preference profile for the shopper. The profile may also include personal data or characteristics (e.g. age, occupation, address, hobbies) regarding the shopper as provided by the shopper when subscribing to the service or obtained from other sources. Profile data can help in discerning the meaning of words used in a keyword query. For instance, a keyword in the query of a medical doctor could have an entirely different meaning to the use of the same keyword presented by a civil engineer. The data accumulation on the shoppers are placed in the shoppers profile database 112 or 118 of each of the merchants. Each individual shopper's profile in the databases of the merchants and the search application service providers can differ from one to another based on the particular merchant's or service providers experience with the shopper and their profiling software. Data collection may continue during searches made by the shopper so that up-to-date profile data for the shopper is obtained and used.

With information regarding the shopper involved in the shopping transaction, the merchant is able to meet the needs of the shopper, and the shopper is presented with the opportunity to view and purchase that merchandise that is most likely to be of interest since the merchant's products and services are directed toward those shoppers who have, either directly or indirectly, expressed an interest in them.

When the search characteristics in the form for key words are entered by the shopper into the space provided on the default or home page of his/her browser, the search engine of the merchant web server 102 does a search of the accessed full text index database 110 or 118 using the key words and gets a list of documents describing those products and services that contain matches to the key words. This list of documents contain basic test ranking Tf (including the number of hits, their location, etc. which are used to order the list of documents) with documents with higher scores at the top. This list is then sent to a ranking module which will apply a ranking algorithm, such as the one described in the article entitled "The Anatomy of a Large-Scale Hypertextual Web Search Engine" by Sergey Brin and Lawrence Page of the Computer Science Department, Stanford University, Stanford CA 94305 (which article is hereby incorporated by reference) to rank the list of documents using the text factors and other rank factors, such as link analysis, popularity, the user's preferences from the users profile, and may also introduce factors reflecting the information, providers biases and interests. A reordered list of documents based on the ranking algorithm is then provided to the user.

Figure 3 shows how a multi-language internet search management server 120 can be used as one of the merchants web server 120 obtain information from the merchant and supply it to a user. As shown in Figure 2, the search management server 120 is connected in a private intranet network 200 with a server 202 and a number of computers 100, such as those described in Figure 1, so that the computers 100 can obtain information stored in the internal sources of the private intranet. The intranet 200 is provided with public internet access capability which provides access to services on the public internet 104. A "firewall" 222 separates the public internet 104 from the private intranet 200 allowing only those with the proper ID and password to enter the intranet 200 from the public internet 104. Internal sources of the intranet 200 are company document management systems 204, and internal databases 206. Also, intranet 200 is provided with a speech recognition system 220 capable of responding to compressed digitized data of voice commands and voice dictation provided by the client computers 100 either from an individual computer 100 or a client's network of such computers.

In the above mentioned U.S. application serial #10/180,195, the search management server 120 contains an integrated search management system which receives queries and information from search engines both in the intranet and internet and accesses information sources other than those that are in the intranet and internet through the computers 100. For example, voice messages transmitted to computer 224 and connected to text by a speech recognition system 220 can be stored in the integrated search management system. The integrated management server contains a central processing unit 230, network interfaces 232 and sufficient random access memory 234 and high density storage 236 to perform its functions. In addition to its connection to the intranet, the search management system contains a direct link 226 to the internet to enable access by customers of the merchant.

Recently, the number of search systems and search engines types grew rapidly. For each given domain, a diversity of specialized search engines could be presented as potential candidates offering different features and performances. While these specialized search systems are invaluable in restricting the scope of searches to pertinent classes, as pointed out above, relevant documents are missed. This is particularly troublesome in technically oriented databases where unsuccessful search queries resemble one another resulting in dissatisfaction. This invention provides a solution to this problem through a search enhancement that involves examination of previous search results received by customers in response to unsuccessful queries. Unsuccessful queries may be ones that return too few references (say less than 5) or ones that have elicited customer complaints.

As shown in Figure 4, a semantic taxonomy 400 provides an input to a semantic binder 402 which binds a semantic category in the semantic taxonomy 400 to the possible query terms appearing in documents of a database by placing the semantic category in the textual index for the documents. For instance, 'video player' 404 is bound to the semantic category "multimedia" 406 in the textual index 408 in each of the documents containing the phrase 'video player'. Similarly, the other of the terms "DVD driver" and "multimedia software" are attached to "multimedia" in the textual indices 410 and 412 in each of the documents. When a search is to be performed in the textual index 408, the search not only uses the original search terms of Figure 1 ('video player') but also searches the semantic "multimedia" for the term. Thus the search would not only turn up documents containing 'video player' but also documents containing the terms "DVD driver" and "multimedia software" bound to the same document semantic ("multimedia").

As shown in Figures 5 and 6, whenever the system database 402 is queried by new query terms, the system accesses the log database 502 (step 600) and places the new terms in a file in the dictionary builder 506 (Step 602). The text analyzer 504 analyzes the query terms in the file (step 604) and engages dictionary builder 506 to associate the query terms in the document with one or more semantic nodes of the semantic taxonomy 400 in the query semantic dictionary 508 (step 606). When the document has been analyzed and no more documents are to be analyzed the process is terminated (step 608).

Referring now to Figures 5 and 7, when a new document, added to the database, is identified (step 703), the documents content is stored in the file system (step 702) and the textual analyzer 504 analyzes the saved file (step 704). Based on this analysis, the semantic binder 510 links query terms in the document 512 by placing the node term ("multimedia") with the query term in the textual index for the document (step 706). When all documents have been processed, the analysis is terminated (step 708). This binding to semantic nodes will cause the semantic nodes to be interrogated together with the link query terms.

Referring now to Figure 8, a user logs onto the system and submits a keyword search (step 801) sending the search query 802 to the search application 804. The search application 804 looks up the query keywords in query semantic dictionary 806 to see whether any corresponding semantic node 807 can be found. If there is (are) matched semantic node(s), search application transforms user's query into["original query OR (semantic nodes)"] (here, "video player for PC" OR "multimedia") so that relevant documents (even though they do not contain any user's query keyword) can also be found in the textual index. The search application then sends the expanded query to search engine 809 and receives search results. The matched documents, including 1) documents that contain user's query keywords 408 and 2) other documents 410 and 412 that belong to the queried semantic node "multimedia", are returned to the user.

The search application places the query term 802 into the log files 810. The text analyzer 812 scans through the log files to the find query keywords processed by the search application 804 and calculates how many times a query has been submitted and records that into the log database, along with the query as described previously. Dictionary builder 816 then updates the listing query semantic dictionary based on the scanning of log files to increase the terms defined in the semantic dictionary 806. The query terms are arranged in the query semantic dictionary in order of "most often queried terms" so that the time of corresponded semantic looking up is short to significantly improve search accuracy with less effort.

In the above figures the textual analyzer can take many forms. It can be simple lookup tables that link the semantic dictionary to the incoming queries or documents. It can also use known commercially available analysis packages for analysis of the query terms. As shown in Figure 9, in more sophisicated systems the transaction analyzer, the semantic binder and the dictionary builder can include submodules similar to those found in Figure 5 of copending application, serial # CHA920030010US1.

The semantic binder can include the following sub-modules:

a sub-module 900 that identifies domain specific terms in a given query, using domain specific glossary 902 relating to the semantic taxonomy.

a sub-module 904 that finds synonyms and related terms for the identified terms, using domain specific thesaurus 906.

a sub-module 908 that finds statistically close terms using listings of associated sets of terms 910.

a sub-module 912 that identifies relevant semantic taxonomy specific categories for the query terms, using domain specific ontology 914.

The dictionary builder 506 can include a sub-module 916 that binds queries in the identified semantic taxonomy categories using the results of the text analyzer 504.

The semantic binder 510 can include a sub-module 915 that adds new doc-query links to the meta-data of the textual index entries to link the documents to the semantic taxonomy categories.

The Index/Meta-data Enhancer module modifies the original Textual Index 524, creating Enhanced Textual Index that replaces the original Textual Index and allows to find more relevant documents in response to the given query.

With this design, the search application not only applies Boolean operations (AND, OR NOT) on end-user's query terms, but also it tries to figure out what the end-user is really looking for based on the info within the query semantic dictionary. For example, search on "pc video player" will also bring back all documents related to "multimedia software that can be executed on a PC."

Above described is one embodiment of the invention. Of course a number of changes can be made. For instance the techniques of the present invention generated in accordance with that copending application can be combined with those of the above mentioned copending application, serial # (CHA920030010US1) filed on even date herewith to select new semantic categories as described in this application. As an example, the "multimedia" category could be divided into subcategories for "computers" and "televisions" as the number of queries relating to "video player" warranted such a distinction. Therefore it should be understood that while only one embodiment of the invention is described, a number of modifications can be made in this embodiment without departing from the spirit and scope of the invention as defined by the attached claims.